Anderson, Christopher

PREDICT 456 Section 55

Assignment #2

Introduction

Analysis performed in Assignment #1 examined the relationship between height and pitching in Major League Baseball. Positive correlations were found between height and both perceived velocity and the downward plane of the ball as it travels towards home plate. However, assessing the impact of these variables on pitching effectiveness revealed inconclusive results. Thus, the question remains: why do professional baseball teams prefer taller pitchers?

As mentioned in the previous report, the theory is that taller pitchers use their size to throw the ball harder and create a greater downward plane when the ball is released (Greenberg, 2016). The difference between tall and short MLB pitchers has been studied before. Baseball analyst David Cameron wrote that teams often undervalue short pitchers due to a belief that their bodies will not hold up to a large workload (Cameron, 2003). Baseball writer Jeff Zimmerman suggests that shorter pitchers receive less credit for their ability and found that they actually outperform others (Zimmerman, 2014). Elliot Evans of the baseball research website Fangraphs.com examined tall and short pitchers from the perspective of performance per dollar and found that short pitchers are undervalued, despite no significant difference in performance (Evans, 2015). An investigation by Driveline Baseball, a training company specializing in the biomechanics of pitching, found that the mechanics of taller pitchers actually creates more stress on the arm than shorter pitchers (Boddy, 2016). While many believe that shorter pitchers may be undervalued, the trend in MLB has not changed.

The purpose of this analysis is to examine the prevalence of tall MLB pitchers from a different perspective, building on the previous analysis by using PITCHf/x data to compare tall pitchers to their shorter counterparts. Created in the early 2000s and used in every MLB stadium since the beginning of the 2008 season, PITCHf/x captures data on the release point, speed, and trajectory of every pitch that occurs in every game throughout the league (Marchi, 2013). The hypothesis for this study is that the data will reveal a significant difference between tall and short pitchers in at least one variable, shedding light on the reason behind MLB's roster selection strategy.

Methods

PITCHf/x results data from the beginning of the 2008 season through the first half of the 2016 season for pitchers who threw a minimum of 50 pitches during the period was extracted from MLB's Statcast database (Statcast Search, 2016). While the dataset used for Assignment #1 contained a record for every pitch thrown during the relevant period, the results data used for this study included one record for each

pitcher who played during the period. The Statcast database summarized the pitch-by-pitch data into key performance metrics that indicate the average velocity, perceived velocity, and spin rate for each pitcher. It also included the average exit velocity and launch angle of all batted balls per pitcher, as well as a few rate measurements like batting average that indicate how well batters perform against each pitcher. Using results data rather than pitch-by-pitch data drastically reduced the size of the dataset, so historical data spanning the entire PITCHf/x era was gathered in order to generate as many observations as possible. A minimum of 50 pitches thrown was used to eliminate position players who occasionally pitch in blowout situations. They are typically untrained in pitching and would likely create outliers in the data. Finally, additional data was downloaded from the Baseball Prospectus Active Player tables to supplement the dataset with the pitcher's height for each record (Baseball Prospectus | Active Players by Year, 2016). Each variable in the dataset is defined below in Table 1.

Table 1: Variable Definitions

Variable	Description	Type	Example
player_id	Unique identifer used by MLB.com	Nominal	110683
player_name	Name of the pitcher	Nominal	Miguel Batista
height	Height of the pitcher (inches)	Ratio	73
pitches	Number of pitches thrown by the pitcher	Ratio	6621
abs	Number of at bats encountered by the pitcher	Ratio	1463
velocity	Average speed of all pitches thrown by the pitcher (MPH)	Ratio	89.12
effective_speed	Average "perceived velocity" of all pitches thrown by the pitcher (MPH)	Ratio	86.22
spin_rate	Average rate of spin on a baseball after it is released by the pitcher (RPM)	Ratio	1924
exit_velocity	Average speed at which the ball leaves the bat against the pitcher (MPH)	Ratio	85.6
launch_angle	Average vertical angle at which the ball leaves the bat against the pitcher	Ratio	7.5
ba	Batting average of hitters against the pitcher	Ratio	0.265
babip	Batting average on balls in play against the pitcher	Ratio	0.302
slg	Slugging percentage against the pitcher	Ratio	0.437
iso	Isolated power against the pitcher	Ratio	0.166

The player_id field is a unique identifier for each player and was used to join the PITCHf/x data to the Baseball Prospectus height data. Velocity, effective_speed, exit_velocity, and launch_angle were used in the previous analysis to investigate the correlation between height and pitching performance. Spin rate is an additional PITCHf/x field that represents the rate of spin on the ball after it is released by the pitcher. The general belief is that pitches with a higher spin rate are more difficult to hit due to the effect on the ball's trajectory (Glossary, 2016). The batting average (ba), batting average on balls in play (babip), slugging percentage (slg), and isolated power (iso) metrics are commonly used today to quantity batter performance. In this case, these fields characterize the performance of all batters combined against each individual pitcher. As the initial step in the exploratory analysis, the descriptive statistics presented in Table 2 were calculated for the ratio level variables.

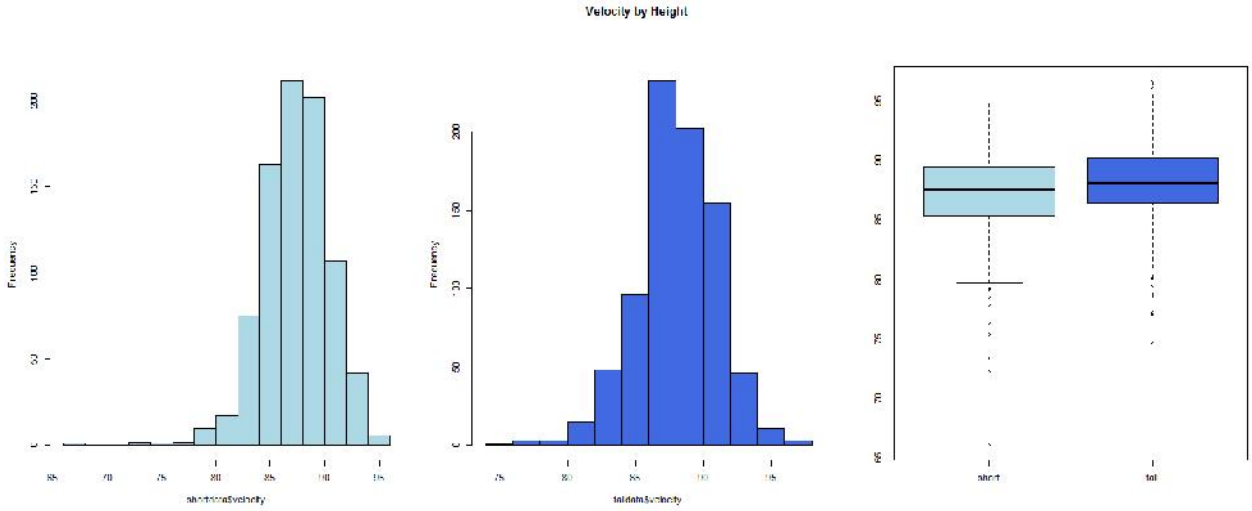
Table 2: Descriptive Statistics

height	pitches	abs	velocity	effective_speed	spin_rate	exit_velocity	launch_angle	ba	babip	slg	iso	tall_short
Min. :66.00	Min. : 50	Min. : 7.0	Min. :66.04	Min. :74.34	Min. : 526	Min. : 0.00	Min. :-2.000	Min. :0.0000	Min. 0.0000	Min. 0.0000	Min. :0.0000	short:839
1st Qu.:73.00	1st Qu.: 404	1st Qu.: 100.0	1st Qu.:85.88	1st Qu.:86.30	1st Qu.:1971	1st Qu.:19.10	1st Qu.: 5.100	1st Qu.:0.2410	1st Qu.:0.2830	1st Qu.:0.3750	1st Qu.:0.1270	tall :814
Median :74.00	Median : 1624	Median : 365.0	Median :87.75	Median :88.20	Median :2096	Median :22.70	Median : 7.000	Median :0.2650	Median :0.3020	Median :0.4190	Median :0.1550	NA
Mean :74.42	Mean : 3669	Mean : 854.6	Mean :87.69	Mean :88.08	Mean :2080	Mean :21.67	Mean : 7.425	Mean :0.2716	Mean :0.3101	Mean :0.4371	Mean :0.1655	NA
3rd Qu.:76.00	3rd Qu.: 4765	3rd Qu.: 1083.0	3rd Qu.:89.78	3rd Qu.:90.27	3rd Qu.:2206	3rd Qu.:23.50	3rd Qu.: 9.100	3rd Qu.:0.2920	3rd Qu.:0.3260	3rd Qu.:0.4720	3rd Qu.:0.1860	NA
Max. :83.00	Max. :23843	Max. :7128.0	Max. :96.62	Max. :97.75	Max. :2567	Max. :95.00	Max. :25.600	Max. :0.6880	Max. :1.0000	Max. :1.2310	Max. :0.7000	NA
NA	NA	NA	NA	NA's :841	NA's :841	NA's :840	NA's :841	NA	NA	NA	NA	NA

In total, there were 1653 observations in the results dataset. Note that there are 841 missing values in the trajectory-related PITCHf/x fields of effective speed, spin rate, exit velocity, and launch angle. There was no explanation for this upon downloading the data, but it was assumed that this was due to missing values in the underlying pitch-by-pitch data used by the Statcast engine to generate the results data (i.e. the PITCHf/x technology failed to capture a pitch). A closer examination of the data revealed that it mostly impacted pitchers who played during the early years of the PITCHf/x era. It is likely that the technology has improved over time and missing values are less common today. The mean height was 74.42 inches, which was used as a threshold to dichotomize the data into two categories and create the new variable tall_short. There were 814 tall pitchers and 839 short pitchers.

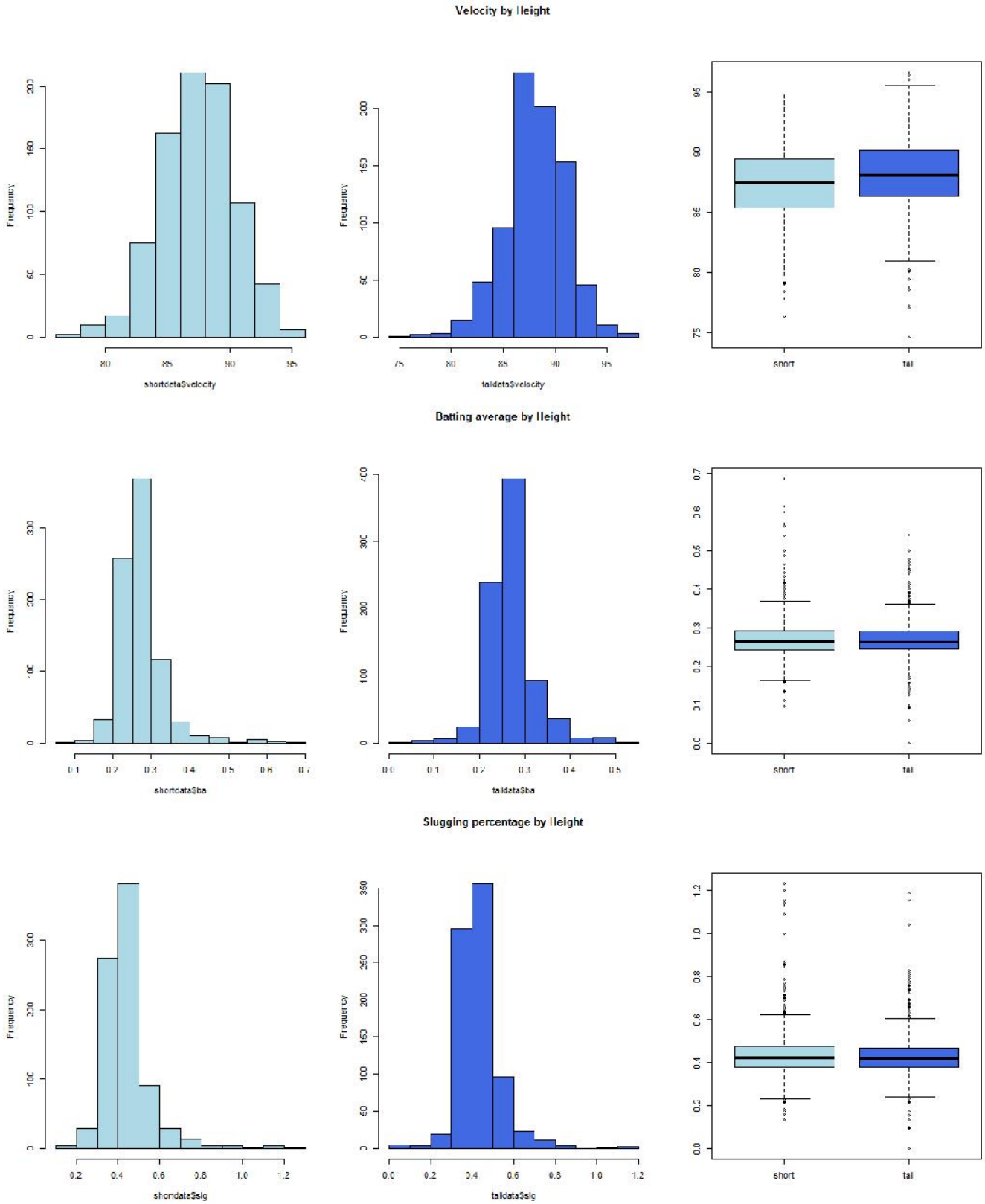
The statistical model used to compare the two categories was Welch's two sample t-test, a variation of the Student's t-test that accounts for differences in variance between the two populations. Otherwise, the assumptions remain the same. The first assumption for the t-test is that the dependent variable(s) should be measured on a continuous scale, which is true in this case. Each item that was examined is a ratio level continuous variable. Next, the independent variable should be categorical with two groups. That is also true in this case, with the dataset split into tall and short pitchers. The third assumption is that there must be independence of observations between the populations or samples, which also holds true for this data. All pitchers are classified into only one group and there is no relationship between the observations in each group. The final two assumptions are that the dependent variable(s) follow a normal distribution for each group and that there are no significant outliers. These assumptions were assessed using histograms and boxplots. Figure 1 on the following page presents the graphics for velocity, with short pitchers presented in light blue and tall pitchers presented in royal blue.

Figure 1: Histograms and boxplots for velocity



There is some obvious negative skew in the distributions for both tall and short pitchers, as well as a few extreme outliers based on the boxplots. One possible explanation of the outliers in the velocity field is knuckleball pitchers. These unconventional pitchers rely almost exclusively on a special pitch called the knuckleball, which is often thrown at a much lower velocity than most pitches. The focus of this study is on the traditional style of pitchers, so known knuckleball pitchers Tim Wakefield, Charlie Zink, Charlie Haeger, Steven Wright, and R.A. Dickey were removed (Wikipedia, 2016). This eliminated 5 observations, leaving a total of 813 tall pitchers and 835 short pitchers (1648 total observations). Figure 2 on the following page presents the plots for a select few variables for discussion purposes. Histograms and boxplots for all variables are included in Figure 4 in Appendix B.

Figure 2: Histograms and boxplots for velocity, ba, and slg



Removing these extreme outliers improved the normality of the velocity distributions, but there are obvious issues with batting average (ba) and slugging percentage (slg) that violate the assumptions of the statistical model. This is also proven by the skewness and kurtosis values presented in Table 3 below.

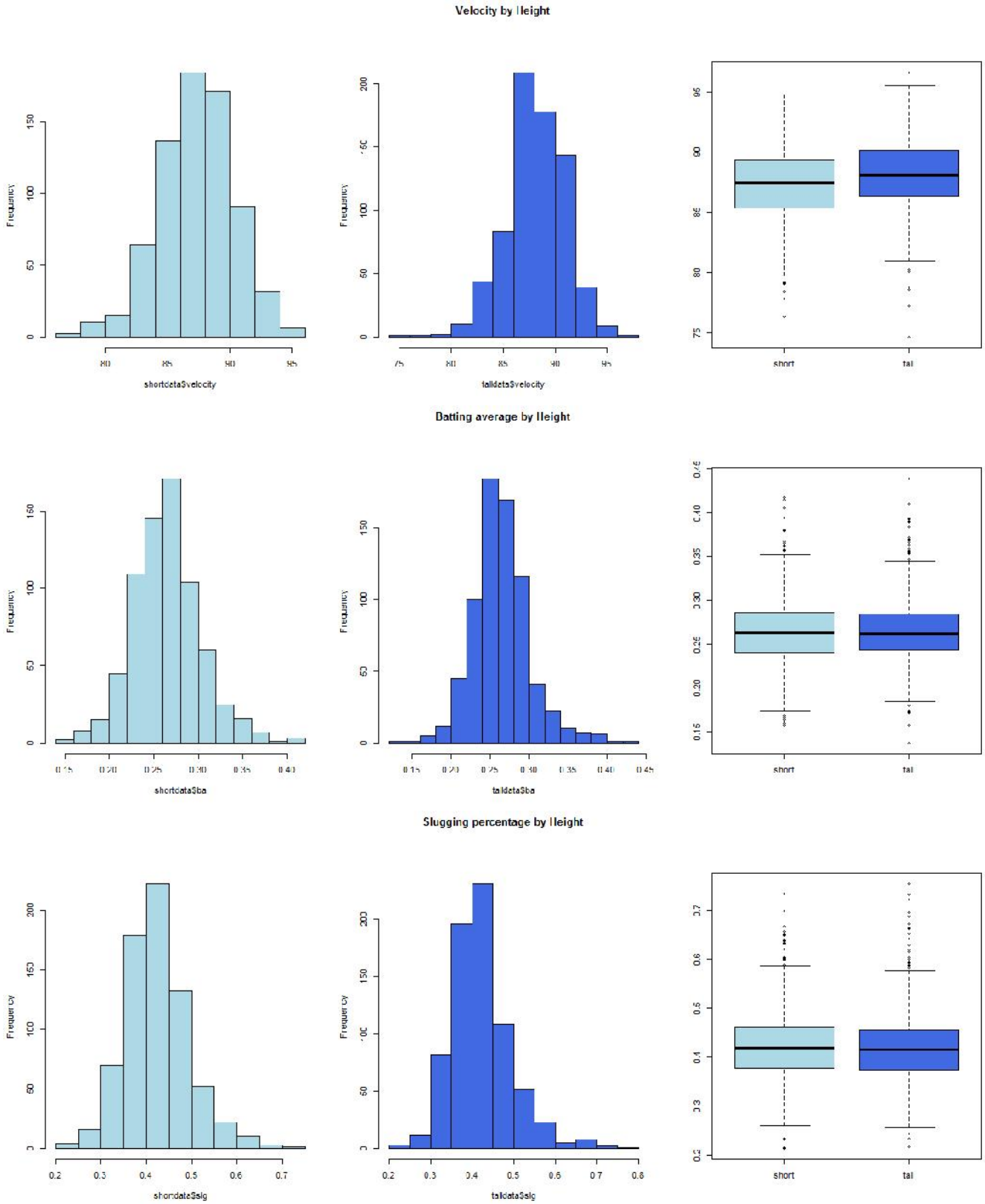
Table 3: Normality assessment

Skewness of 0 and Kurtosis of 3 = Normal Distribution				
	Short pitchers		Tall pitchers	
Variable	Skewness	Kurtosis	Skewness	Kurtosis
velocity	-0.19	3.08	-0.40	3.88
effective_speed	-0.27	3.05	-0.36	3.57
spin_rate	-0.30	3.26	-0.28	3.28
exit_velocity	1.86	4.88	1.85	4.88
launch_angle	1.29	5.89	0.99	5.96
ba	2.17	12.71	0.75	7.65
babip	3.05	26.90	0.47	9.26
slg	2.54	14.34	1.71	12.32
iso	2.63	15.14	2.06	13.66

There are a few variables that approach normality in both the tall and short groups. However, there are also several variables that show drastic deviation from normality, particularly the batting performance metrics batting average (ba), batting average on balls in play (babip), slugging percentage (slg), and isolated power (iso).

In addition to the apparent normality violations, the boxplots for batting average (ba), slugging percentage (slg), and several other variables show a large number of outliers. By examining the dataset in a spreadsheet view, it was discovered that many of the extreme values for these variables that lie near the maximum and minimum ends of range occur in observations where the number of at bats was relatively small. For these particular pitchers, it appeared that the sample size was not large enough to deliver results that are indicative of the typical MLB style of play. Therefore 216 observations with less than 50 at bats were removed from the dataset. Consequently, 721 tall pitchers and 711 short pitchers remained, for a total of 1432 observations. At this point, normality was checked again on the modified data. Figure 3 below presents the plots for a select few variables after outlier removal for discussion purposes. Histograms and boxplots for all variables are included in Figure 5 in Appendix B.

Figure 3: Histograms and boxplots for velocity, ba, and slg after outlier removal



A significant improvement in normality is obvious in the histograms for batting average (ba) and slugging percentage (slg). While the boxplots continue to show outliers, it appears that the volume has decreased significantly.

Table 4: Normality assessment after outlier removal

Skewness of 0 and Kurtosis of 3 = Normal Distribution				
	Short pitchers		Tall pitchers	
Variable	Skewness	Kurtosis	Skewness	Kurtosis
velocity	-0.22	3.15	-0.39	3.77
effective_speed	-0.26	3.15	-0.35	3.53
spin_rate	-0.32	3.42	-0.29	3.37
exit_velocity	2.45	7.86	2.12	6.12
launch_angle	0.94	5.11	0.95	6.37
ba	0.40	3.97	0.68	4.96
babip	0.50	5.15	0.81	5.91
slg	0.60	4.28	0.94	5.01
iso	0.83	4.99	1.17	5.47

This improvement extends to other variables, as well, and the skewness and kurtosis values presented in Table 4. Many of the distributions for both tall and short pitchers exhibit some degree skewness and kurtosis. However, none of these values represent significant violations of the normality assumption, particularly when combined with the large sample size of over 700 observations in each group. After the removal of outliers based on unconventional styles and a minimum number of at bats, the assumptions for performing the t-tests have been met. Notably, the distributions for exit velocity diverge from normality more than any of the others. A non-parametric test may be necessary to properly evaluate this metric.

Results

The results of the individuals Welch's t-tests are displayed in Table 5 below, with the statistically significant p-values highlighted in yellow.

Table 5: Results of Welch Two Sample t-tests

Variable	Mean (Short)	Mean (Tall)	p-value	95% CI
velocity	87.34	88.13	4.01E-07	-1.09 to -0.48
effective_speed	87.59	88.51	4.07E-05	-1.37 to -0.49
spin_rate	2073.75	2095.32	0.09	-46.78 to 3.63
exit_velocity	28.58	30.15	0.30	-4.53 to 1.40
launch_angle	7.27	7.25	0.95	-0.48 to 0.51
ba	0.265	0.265	0.93	-0.0037 to 0.0041
babip	0.303	0.305	0.26	-0.0061 to 0.0016
slg	0.424	0.422	0.68	-0.0059 to 0.0091
iso	0.159	0.157	0.57	-0.0035 to 0.0063

The small p-values for velocity and effective speed denote a statistically significant difference in the mean values of each group. The mean value for both variables is higher in the tall group than in the short group. Therefore, this can be interpreted to mean that the tall pitchers in this dataset throw the ball harder than the short pitchers. The p-value for spin rate is not significant at the 95% confidence level, but would be significant at a 90% confidence level. The mean value is higher for tall pitchers, suggesting that they may achieve a greater spin rate than shorter pitchers. However, this conclusion does not share the same degree of certainty as seen with velocity and effective speed. All of the other variables exhibit p-values that are demonstrably insignificant at a 95% confidence level. From these results, it can be inferred that there is no difference between tall and short pitchers in regards to exit velocity, launch angle, batting average (ba), batting average on balls in play (babip), slugging percentage (slg), and isolated power (iso).

For the sake of thoroughness, the non-parametric Wilcoxon Rank Sum Test was also performed on each pair. The results were identical to those shown above, with statistically significant differences present on only velocity and effective speed. It was observed that a non-parametric test may be more appropriate for exit velocity, but both the Wilcoxon and the Welch test indicate that a difference between tall and short pitchers is not likely.

Implications

The hypothesis that the data will reveal a significant difference between tall and short pitchers in at least one variable was true. Both velocity and effective speed presented a significant difference between tall and short pitchers at the 95% confidence level. In both cases, the mean value for taller pitchers was larger so it was concluded that taller pitchers do, in fact, throw harder than shorter pitchers. Furthermore, spin rate showed a significant difference between tall and short pitchers at the 90% confidence level. The mean spin rate for the tall group was higher than the mean value for the short group. As mentioned in the introduction, a greater spin rate is associated with a greater degree of difficulty for the batter. No difference was found in the metrics that indicate how batters fare against tall or short pitchers (e.g. batting average or slugging percentage).

The findings presented above imply that MLB teams prefer taller pitchers largely because of their ability to throw the ball harder than shorter pitchers. The tendency does not appear to be based on any performance because there was no difference between tall and short pitchers in opponents' batting average, BABIP, slugging percentage, or isolated power. This offers an explanation for roster construction throughout the league that focuses on height. Also, in an era where the sabermetrics movement has provided more performance measuring statistics than ever before, the results suggest that teams still seek the raw skills of bigger, stronger athletes who can throw with greater velocity. Teams may be able to find market inefficiencies in shorter pitchers who achieve better results than taller pitchers who may throw the ball harder.

One limitation of the data was discussed in the Methods section in that the effective speed, spin rate, exit velocity, and launch angle were missing for over 800 pitchers. It was assumed that these missing

values were caused by technological limitations in the early years of the PITCHf/x system. Also, this study did not distinguish between starting pitchers and relief pitchers. Often, these two types of pitchers have very different styles and goals and it may be wise to examine these relationships in that context. Likewise, the data could be further partitioned into left-handed and right-handed throwers. In addition, a value other than the overall mean height could have been used to distinguish between tall and short pitchers. This cutoff value was 74.42 inches, or roughly 6 feet 2 inches. While this may be the mean height of MPB pitchers, it is still much taller than the average human male, so many pitchers in the short group would be considered tall outside of baseball. Other pitching metrics, such as ERA or strikeouts, could have been included as well to look at performance.

An alternative to the t-tests performed in this analysis would have been an ANOVA model, which examines all dependent variables at once and indicates whether there is a difference in any of the variables. However, the t-test model was chosen because the tall and short populations do not contain an equal number of observations. In addition, the individual t-tests offer a more granular breakdown of the differences between tall and short pitchers that will hopefully identify the aspects of pitching on which taller pitchers perform better.

Appendix A: References

Baseball Prospectus | Active Players by Year. (2016). Retrieved July 10, 2016, from http://www.baseballprospectus.com/sortable/extras/active_players.php

Baseball Prospectus | Glossary. (2016). Retrieved July 10, 2016, from http://www.baseballprospectus.com/glossary/

Boddy, K. (2016, June 16). Tall Pitchers vs. Short Pitchers – Velocity, Elbow Injuries, and Mechanics. Retrieved July 31, 2016, from https://www.drivelinebaseball.com/2016/06/16/tall-pitchers-vs-short-pitchers-velocity-elbow-injuries-mechanics/

Cameron, D. (2003, July 3). Prospecting: Short Pitchers. Retrieved July 31, 2016, from http://www.baseballprospectus.com/article.php?articleid=2064

Evans, E. (2015, April 3). Six Feet Under: Evaluating Short Pitchers. Retrieved July 31, 2016, from http://www.fangraphs.com/community/six-feet-under-evaluating-short-pitchers/

Glossary. (2016). Retrieved July 10, 2016, from http://m.mlb.com/glossary/statcast/

Greenberg, G. P. (2010). Does a Pitcher's Height Matter? Retrieved July 10, 2016, from http://sabr.org/research/does-pitcher-s-height-matter

List of knuckleball pitchers. (2016, July 27). Retrieved July 31, 2016, from https://en.wikipedia.org/wiki/List_of_knuckleball_pitchers

Marchi, M., & Albert, J. (2013). Analyzing baseball data with R. Boca Raton, FL: CRC Press.

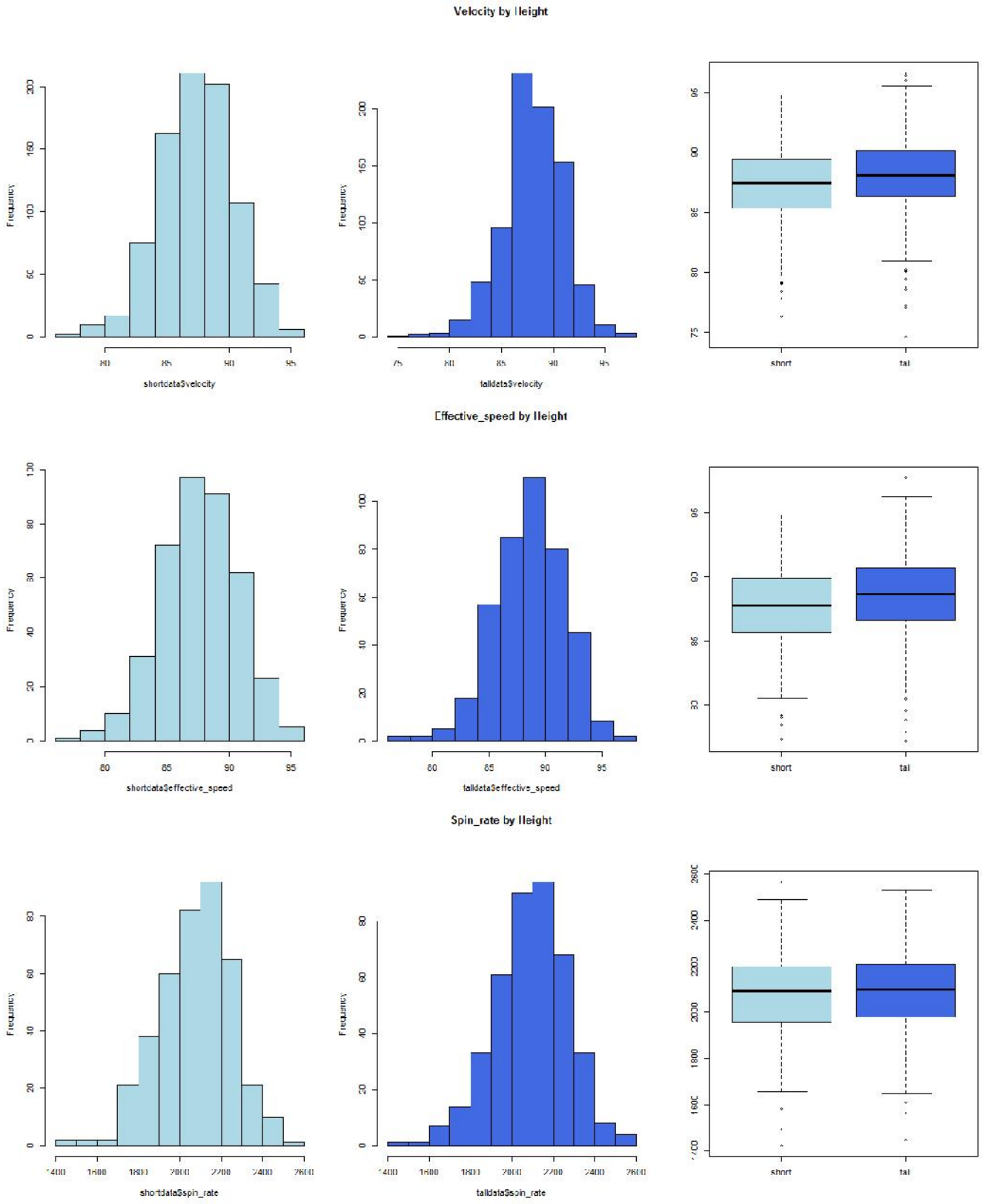
Rymer, Z. D. (2013, May 13). Do Taller Pitchers Throw Harder Than Average? Retrieved July 10, 2016, from http://bleacherreport.com/articles/1645950-do-taller-pitchers-throw-harder-than-average

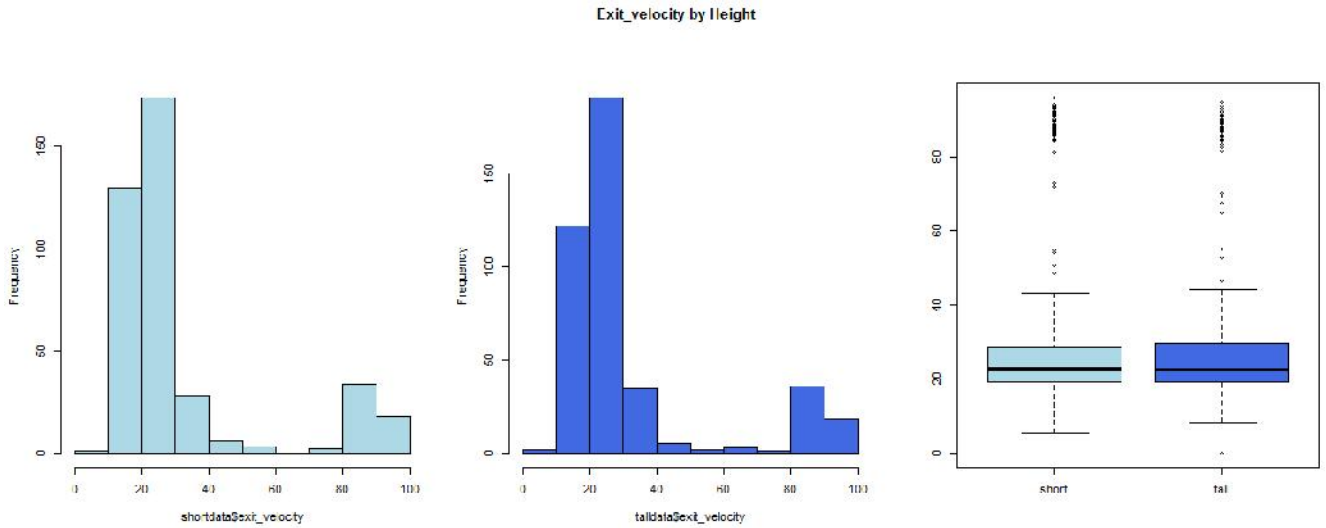
Statcast Search. (2016). Retrieved July 10, 2016, from https://baseballsavant.mlb.com/statcast_search

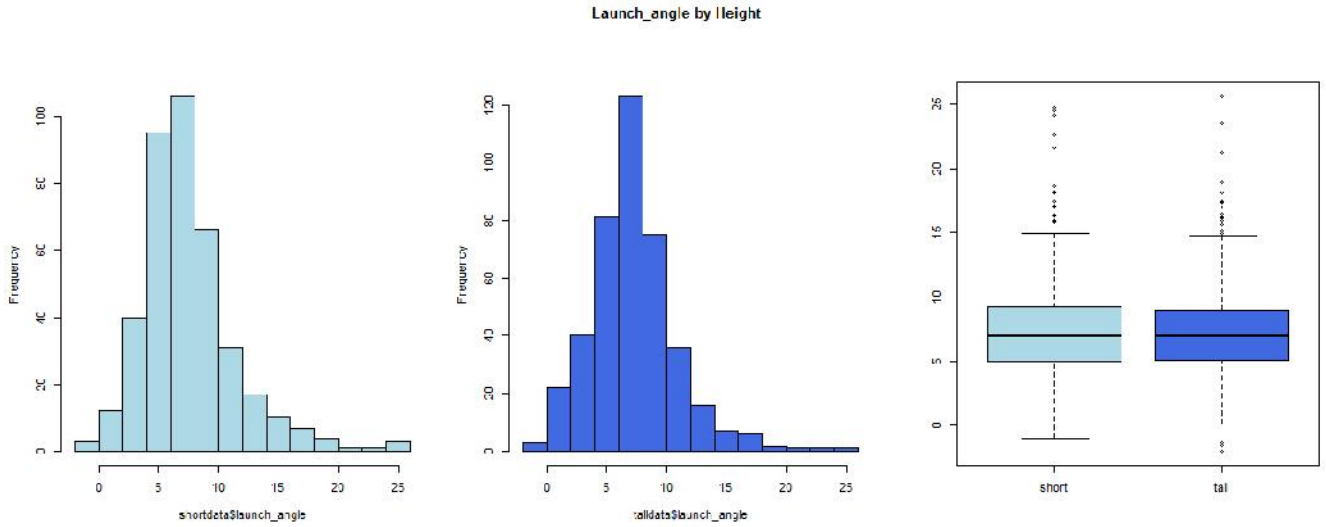
Zimmerman, J. (2014, October 6). Should Short Pitchers Still Get Short Shrift? Retrieved July 31, 2016, from http://www.hardballtimes.com/short-pitchers-still-getting-short-shrift/

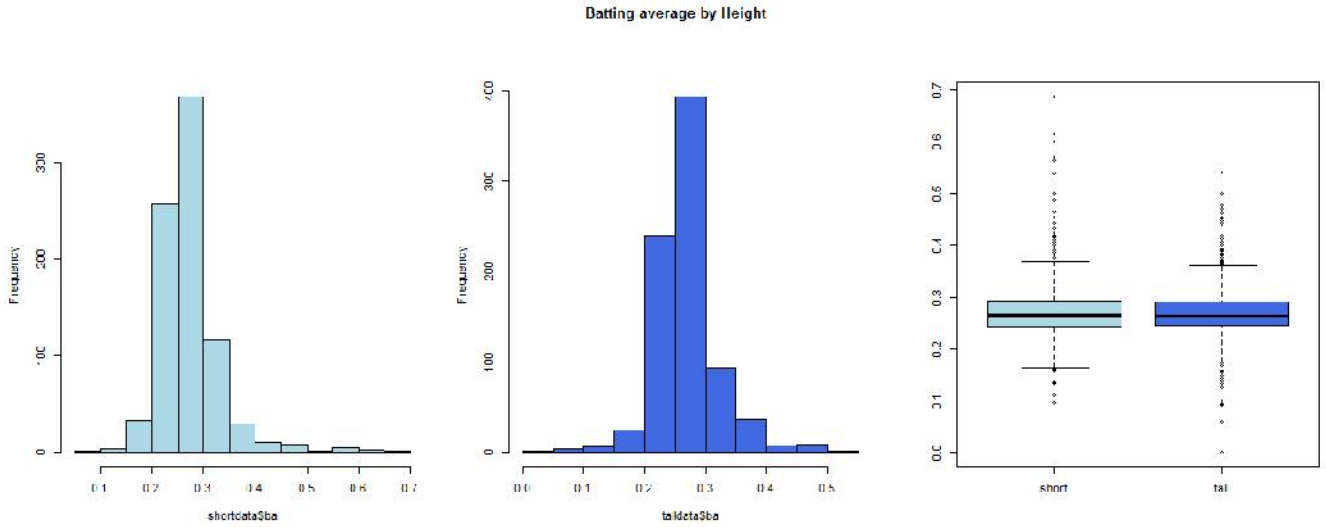
Appendix B: Expanded Figures

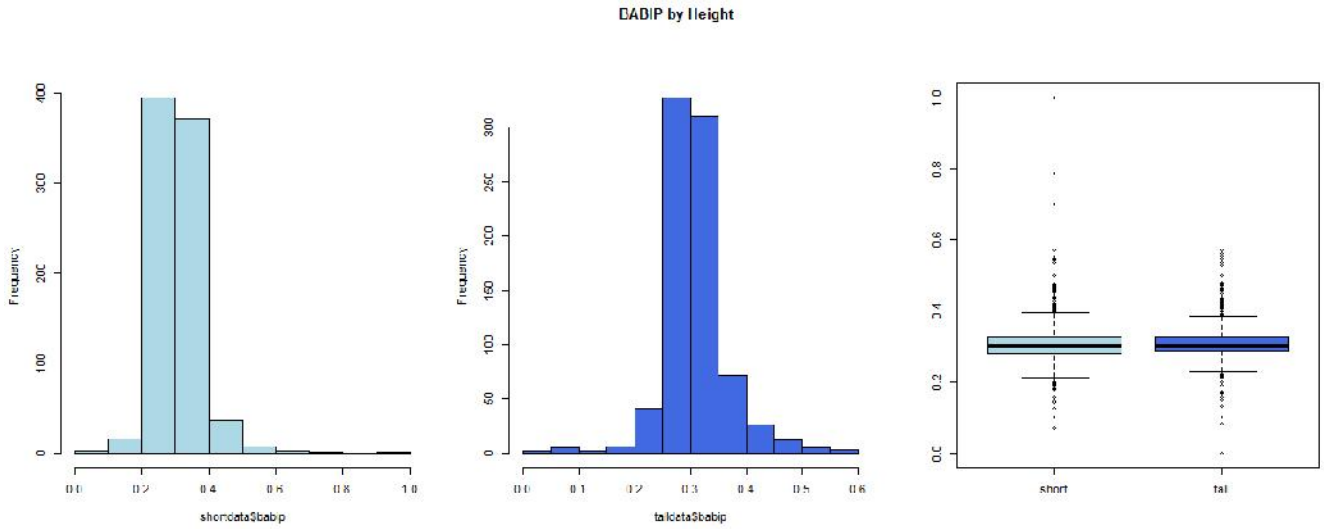
Figure 4: All histograms and boxplots

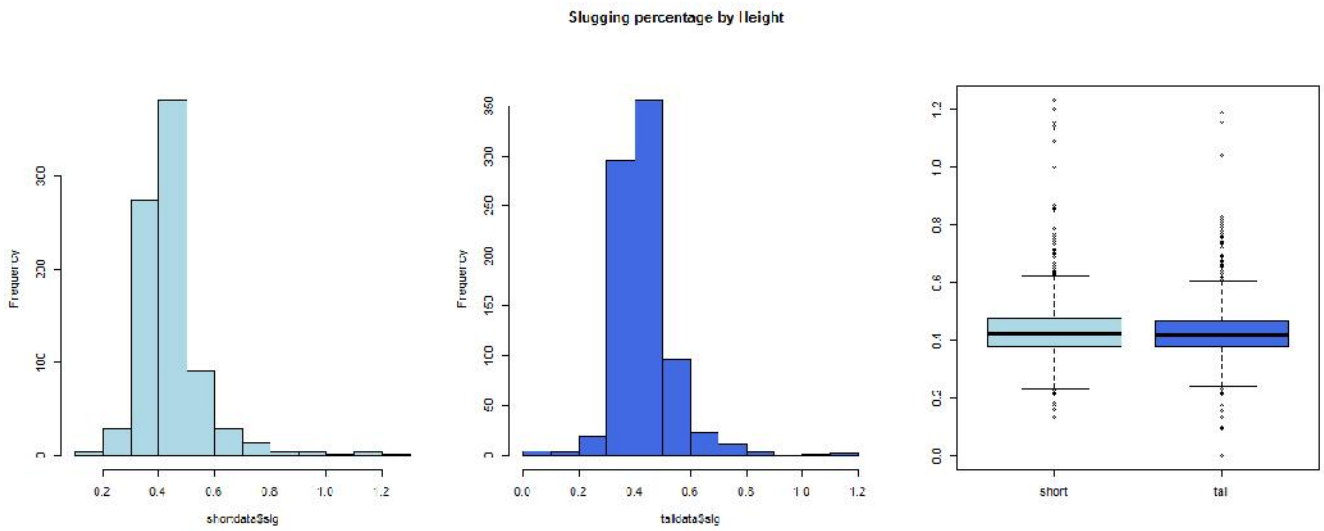



Exit_velocity by Height
Frequency
shortdata$exit_velocity
talldata$exit_velocity
short
tall


Launch_angle by Height
Frequency
shortdata$launch_angle
talldata$launch_angle
short
tall


Batting average by Height
Frequency
shortdata$ba
talldata$ba
short
tall


BABIP by Height
Frequency
shordata$babip
talldata$babip
short
tall


Slugging percentage by Height
Frequency
shordata$slg
talldata$slg
short
tall

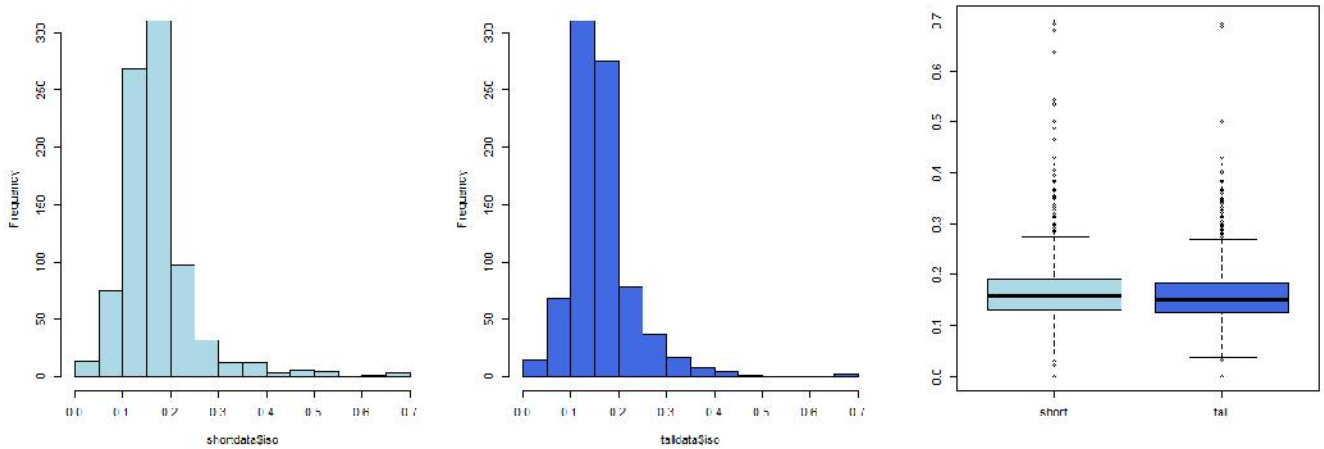
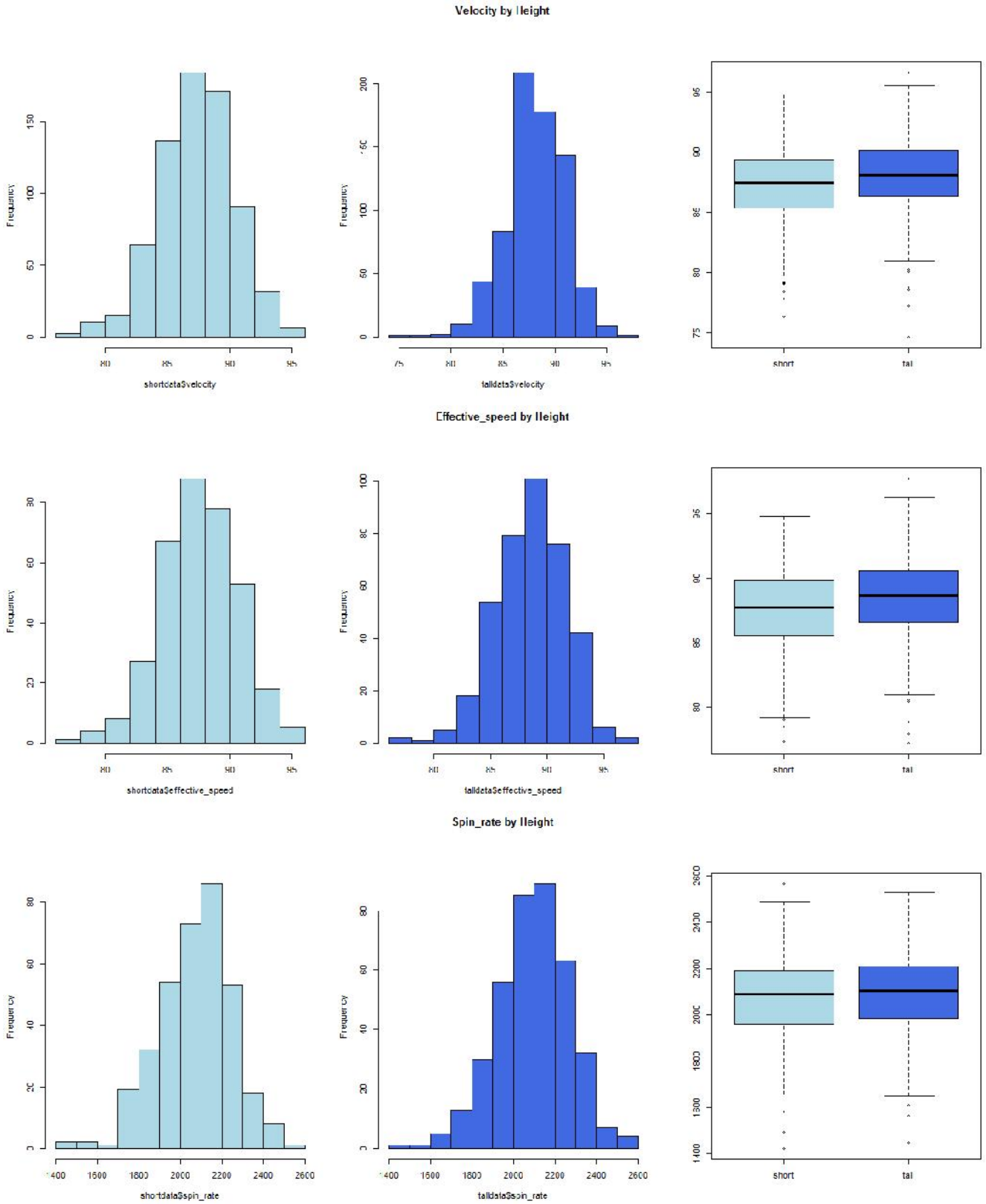
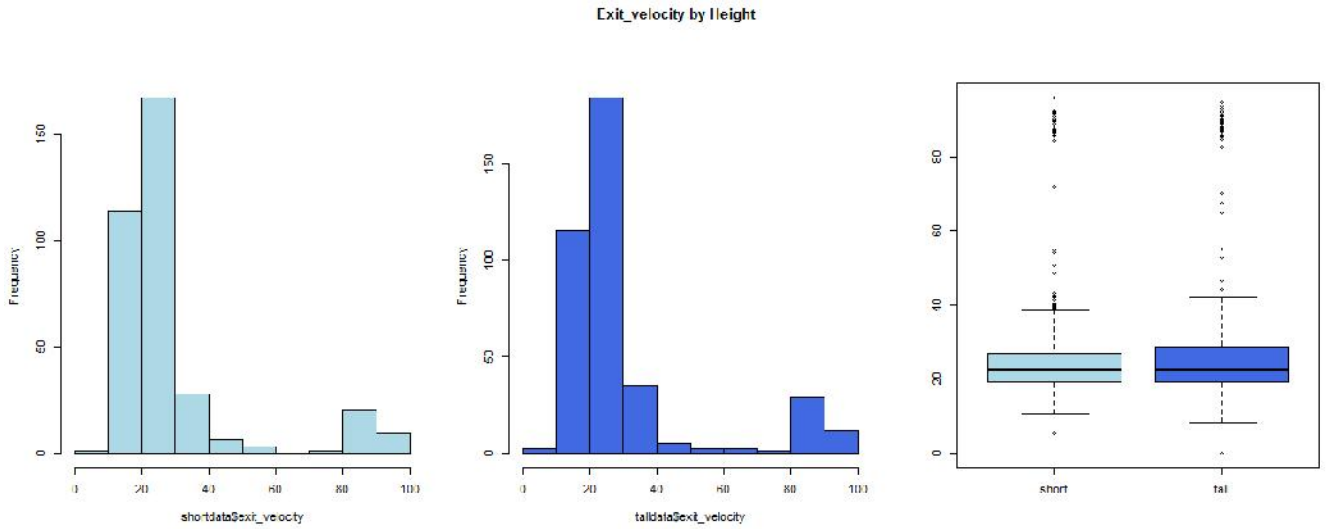
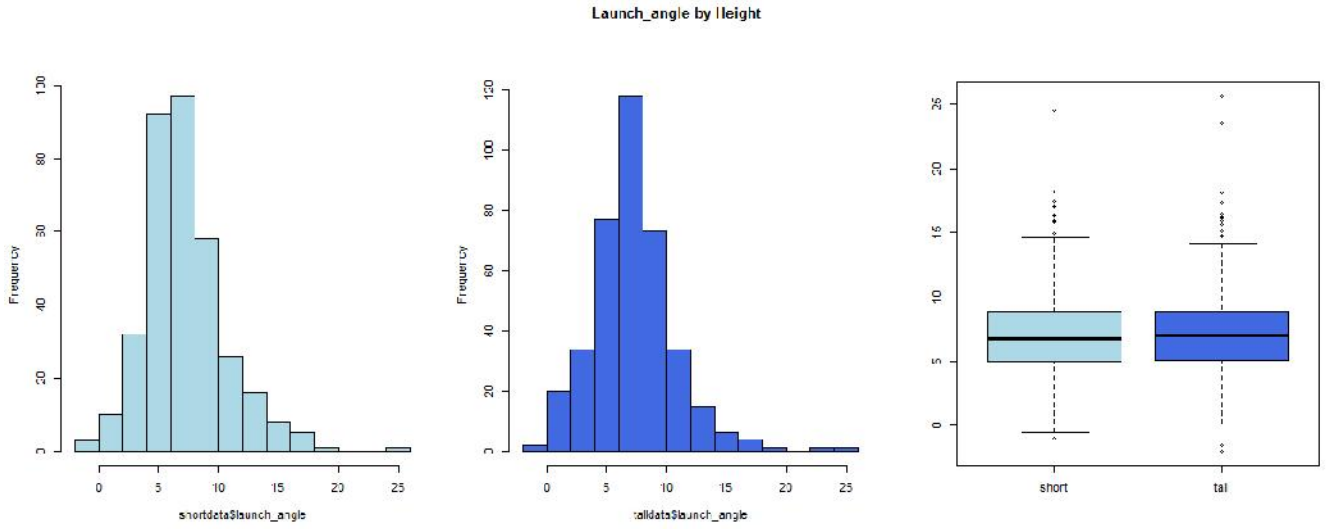
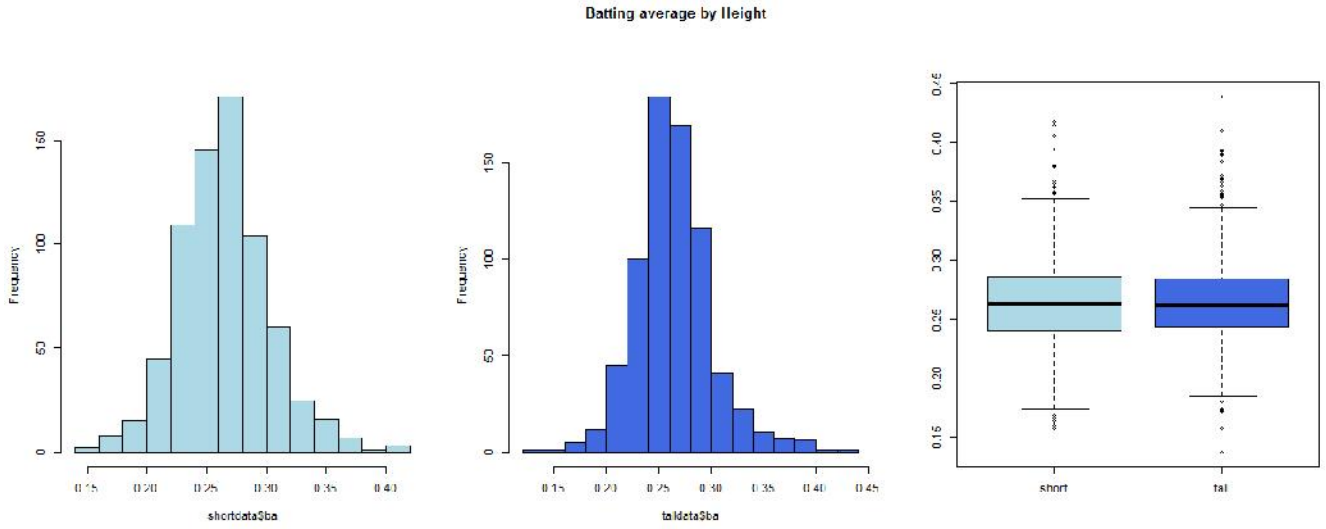

Isolated power by Height
Frequency
shordata$iso
talldata$iso
short
tall

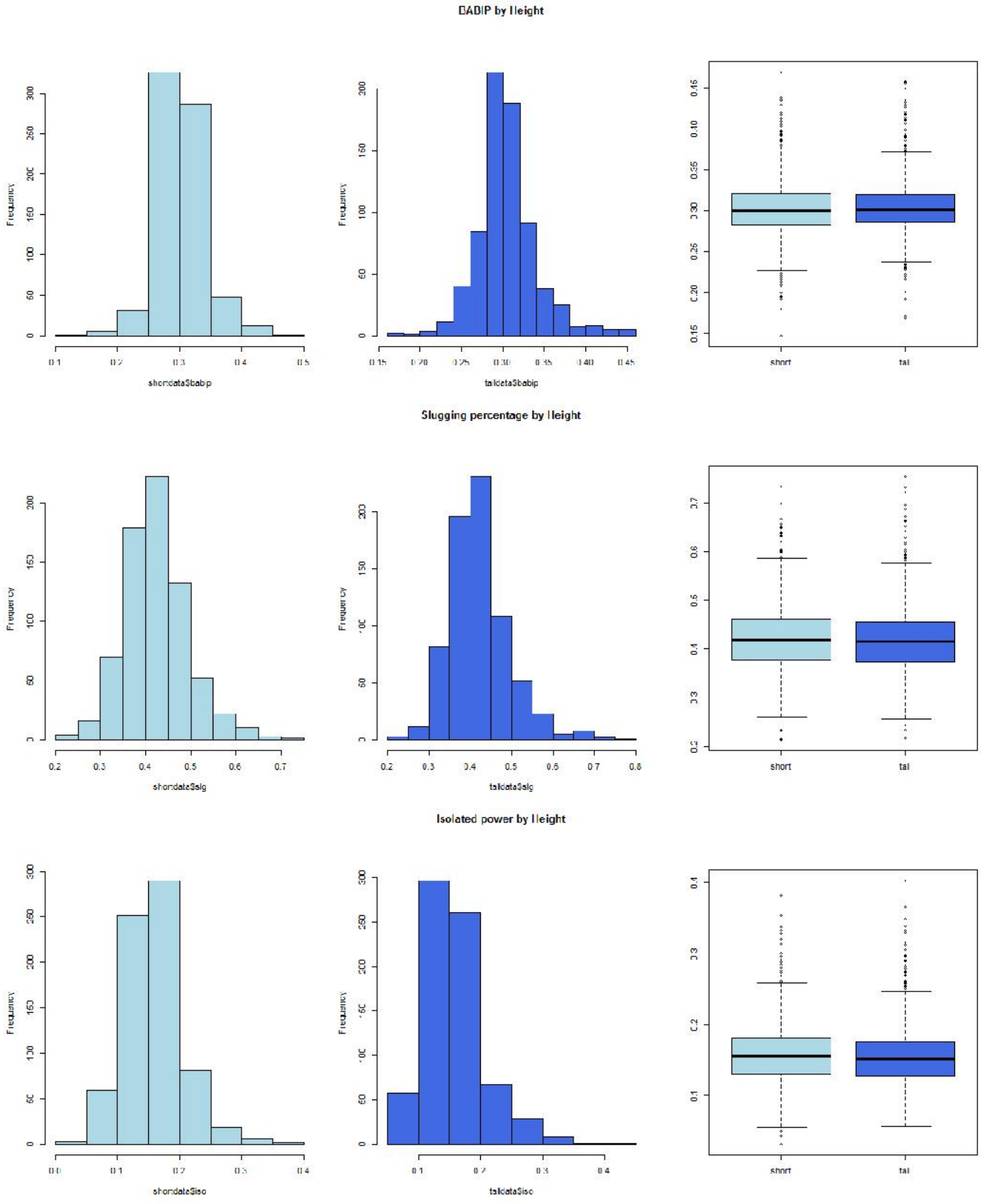
Figure 5: All histograms and boxplots after outlier removal




Exit_velocity by Height
Frequency
shortdata$exit_velocity
talldata$exit_velocity
short
tall


Launch_angle by Height
Frequency
shortdata$launch_angle
talldata$launch_angle
short
tall


Batting average by Height
Frequency
shortdata$ba
talldata$ba
short
tall


BABIP by Height
Frequency
shordata$babip
talldata$babip
short
tall
Slugging percentage by Height
Frequency
shordata$slg
talldata$slg
short
tall
Isolated power by Height
Frequency
shordata$iso
talldata$iso
short
tall

Appendix C: R Code

```
#PREDICT 456 Sports Performance Analysis Section 55 Summer 2016
#Christopher Anderson
#Assignment #2

library(moments)
library(ggplot2)
library(gridExtra)
library(Hmisc)
library(dplyr)
library(XML)

# Read in pitching results data downloaded from MLB Statcast
results <- read.csv("results.csv", header = T, sep = ",")

# Download and merge player height data from Baseball Prospectus active player
tables
d <-
readHTMLTable("http://www.baseballprospectus.com/sortable/extras/active_players.php
?this_year=2008&this_lvl=MLB")
height2008 <- data.frame(d$pitchers_list_datagrid$`MLB ID`,
d$pitchers_list_datagrid$Height, stringsAsFactors = FALSE)
d <-
readHTMLTable("http://www.baseballprospectus.com/sortable/extras/active_players.php
?this_year=2009&this_lvl=MLB")
height2009 <- data.frame(d$pitchers_list_datagrid$`MLB ID`,
d$pitchers_list_datagrid$Height, stringsAsFactors = FALSE)
d <-
readHTMLTable("http://www.baseballprospectus.com/sortable/extras/active_players.php
?this_year=2010&this_lvl=MLB")
height2010 <- data.frame(d$pitchers_list_datagrid$`MLB ID`,
d$pitchers_list_datagrid$Height, stringsAsFactors = FALSE)
d <-
readHTMLTable("http://www.baseballprospectus.com/sortable/extras/active_players.php
?this_year=2011&this_lvl=MLB")
height2011 <- data.frame(d$pitchers_list_datagrid$`MLB ID`,
d$pitchers_list_datagrid$Height, stringsAsFactors = FALSE)
d <-
readHTMLTable("http://www.baseballprospectus.com/sortable/extras/active_players.php
?this_year=2012&this_lvl=MLB")
height2012 <- data.frame(d$pitchers_list_datagrid$`MLB ID`,
d$pitchers_list_datagrid$Height, stringsAsFactors = FALSE)
d <-
readHTMLTable("http://www.baseballprospectus.com/sortable/extras/active_players.php
?this_year=2013&this_lvl=MLB")
height2013 <- data.frame(d$pitchers_list_datagrid$`MLB ID`,
d$pitchers_list_datagrid$Height, stringsAsFactors = FALSE)
```

```
d <-
readHTMLTable("http://www.baseballprospectus.com/sortable/extras/active_players.php
?this_year=2014&this_lvl=MLB")
height2014 <- data.frame(d$pitchers_list_datagrid$`MLB ID`,
d$pitchers_list_datagrid$Height, stringsAsFactors = FALSE)
d <-
readHTMLTable("http://www.baseballprospectus.com/sortable/extras/active_players.php
?this_year=2015&this_lvl=MLB")
height2015 <- data.frame(d$pitchers_list_datagrid$`MLB ID`,
d$pitchers_list_datagrid$Height, stringsAsFactors = FALSE)
d <-
readHTMLTable("http://www.baseballprospectus.com/sortable/extras/active_players.php
?this_year=2016&this_lvl=MLB")
height2016 <- data.frame(d$pitchers_list_datagrid$`MLB ID`,
d$pitchers_list_datagrid$Height, stringsAsFactors = FALSE)

height <- rbind(height2008, height2009, height2010, height2011, height2012,
height2013, height2014, height2015, height2016)
rm(list =
c("d","height2008","height2009","height2010","height2011","height2012","height2013"
,"height2014","height2015","height2016"))
height <- unique(height)
colnames(height) <- c("player_id", "height")

mydata <- merge(x = results, y = height, by = "player_id")
mydata <- subset(mydata,
select=c("player_id","player_name","height","pitches","abs","velocity","effective_s
peed","spin_rate",

"exit_velocity","launch_angle","ba","babip","slg","iso"))
mydata$height <- as.numeric(as.character(mydata$height))
mydata$player_name <- as.character(mydata$player_name)

# Dichotimize the data into tall and short pitchers
tall_short <- factor(mydata$height > mean(mydata$height), labels = c("short",
"tall"))
mydata <- data.frame(mydata, tall_short)
talldata <- filter(mydata, tall_short == "tall")
shortdata <- filter(mydata, tall_short == "short")
# 1653 total observations 814 tall 839 short

# Examine structure of data and summary statistics
str(mydata)
head(mydata)
tail(mydata)
grid.table(summary(mydata[,3:15]))
dev.off()

# Exploratory histograms and normality assessment
```

```
par(mfrow = c(1,3), oma=c(0,0,2,0))
hist(shortdata$velocity, main = "", col = "light blue")
hist(talldata$velocity, main = "", col = "royal blue")
boxplot(velocity ~ tall_short, mydata, main = "", col = c("light blue","royal
blue"))
title("Velocity by Height", outer=TRUE)
skewness(shortdata$velocity)
kurtosis(shortdata$velocity)
skewness(talldata$velocity)
kurtosis(talldata$velocity)

# Filter out extreme outliers based on known knuckleball pitchers
mydata <- filter(mydata, player_name != "Tim Wakefield")
mydata <- filter(mydata, player_name != "Charlie Zink")
mydata <- filter(mydata, player_name != "Charlie Haeger")
mydata <- filter(mydata, player_name != "Steven Wright")
mydata <- filter(mydata, player_name != "R.A. Dickey")
talldata <- filter(mydata, tall_short == "tall")
shortdata <- filter(mydata, tall_short == "short")
# 5 observations removed; 1648 remaining 813 tall 835 short

# Continue exploratory histograms and normality assessment
par(mfrow = c(1,3), oma=c(0,0,2,0))
hist(shortdata$velocity, main = "", col = "light blue")
hist(talldata$velocity, main = "", col = "royal blue")
boxplot(velocity ~ tall_short, mydata, main = "", col = c("light blue","royal
blue"))
title("Velocity by Height", outer=TRUE)
skewness(shortdata$velocity)
kurtosis(shortdata$velocity)
skewness(talldata$velocity)
kurtosis(talldata$velocity)

hist(shortdata$effective_speed, main = "", col = "light blue")
hist(talldata$effective_speed, main = "", col = "royal blue")
boxplot(effective_speed ~ tall_short, mydata, main = "", col = c("light
blue","royal blue"))
title("Effective_speed by Height", outer=TRUE)
skewness(shortdata$effective_speed, na.rm = TRUE)
kurtosis(shortdata$effective_speed, na.rm = TRUE)
skewness(talldata$effective_speed, na.rm = TRUE)
kurtosis(talldata$effective_speed, na.rm = TRUE)

hist(shortdata$spin_rate, main = "", col = "light blue")
hist(talldata$spin_rate, main = "", col = "royal blue")
boxplot(spin_rate ~ tall_short, mydata, main = "", col = c("light blue","royal
blue"))
title("Spin_rate by Height", outer=TRUE)
skewness(shortdata$spin_rate, na.rm = TRUE)
```

```
kurtosis(shortdata$spin_rate, na.rm = TRUE)
skewness(talldata$spin_rate, na.rm = TRUE)
kurtosis(talldata$spin_rate, na.rm = TRUE)

hist(shortdata$exit_velocity, main = "", col = "light blue")
hist(talldata$exit_velocity, main = "", col = "royal blue")
boxplot(exit_velocity ~ tall_short, mydata, main = "", col = c("light blue","royal
blue"))
title("Exit_velocity by Height", outer=TRUE)
skewness(shortdata$exit_velocity, na.rm = TRUE)
kurtosis(shortdata$exit_velocity, na.rm = TRUE)
skewness(talldata$exit_velocity, na.rm = TRUE)
kurtosis(talldata$exit_velocity, na.rm = TRUE)

hist(shortdata$launch_angle, main = "", col = "light blue")
hist(talldata$launch_angle, main = "", col = "royal blue")
boxplot(launch_angle ~ tall_short, mydata, main = "", col = c("light blue","royal
blue"))
title("Launch_angle by Height", outer=TRUE)
skewness(shortdata$launch_angle, na.rm = TRUE)
kurtosis(shortdata$launch_angle, na.rm = TRUE)
skewness(talldata$launch_angle, na.rm = TRUE)
kurtosis(talldata$launch_angle, na.rm = TRUE)

hist(shortdata$ba, main = "", col = "light blue")
hist(talldata$ba, main = "", col = "royal blue")
boxplot(ba ~ tall_short, mydata, main = "", col = c("light blue","royal blue"))
title("Batting average by Height", outer=TRUE)
skewness(shortdata$ba)
kurtosis(shortdata$ba)
skewness(talldata$ba)
kurtosis(talldata$ba)

hist(shortdata$babip, main = "", col = "light blue")
hist(talldata$babip, main = "", col = "royal blue")
boxplot(babip ~ tall_short, mydata, main = "", col = c("light blue","royal blue"))
title("BABIP by Height", outer=TRUE)
skewness(shortdata$babip)
kurtosis(shortdata$babip)
skewness(talldata$babip)
kurtosis(talldata$babip)

hist(shortdata$slg, main = "", col = "light blue")
hist(talldata$slg, main = "", col = "royal blue")
boxplot(slg ~ tall_short, mydata, main = "", col = c("light blue","royal blue"))
title("Slugging percentage by Height", outer=TRUE)
skewness(shortdata$slg)
kurtosis(shortdata$slg)
```

```
skewness(talldata$slg)
kurtosis(talldata$slg)

hist(shortdata$iso, main = "", col = "light blue")
hist(talldata$iso, main = "", col = "royal blue")
boxplot(iso ~ tall_short, mydata, main = "", col = c("light blue","royal blue"))
title("Isolated power by Height", outer=TRUE)
skewness(shortdata$iso)
kurtosis(shortdata$iso)
skewness(talldata$iso)
kurtosis(talldata$iso)

# Filter out outliers based on minimum number of 50 at bats
mydata <- filter(mydata, abs > 50)
talldata <- filter(mydata, tall_short == "tall")
shortdata <- filter(mydata, tall_short == "short")
# 216 observations removed; 1432 remaining 721 tall 711 short

# Check exploratory histograms and normality assessment again
hist(shortdata$velocity, main = "", col = "light blue")
hist(talldata$velocity, main = "", col = "royal blue")
boxplot(velocity ~ tall_short, mydata, main = "", col = c("light blue","royal
blue"))
title("Velocity by Height", outer=TRUE)
skewness(shortdata$velocity)
kurtosis(shortdata$velocity)
skewness(talldata$velocity)
kurtosis(talldata$velocity)

hist(shortdata$effective_speed, main = "", col = "light blue")
hist(talldata$effective_speed, main = "", col = "royal blue")
boxplot(effective_speed ~ tall_short, mydata, main = "", col = c("light
blue","royal blue"))
title("Effective_speed by Height", outer=TRUE)
skewness(shortdata$effective_speed, na.rm = TRUE)
kurtosis(shortdata$effective_speed, na.rm = TRUE)
skewness(talldata$effective_speed, na.rm = TRUE)
kurtosis(talldata$effective_speed, na.rm = TRUE)

hist(shortdata$spin_rate, main = "", col = "light blue")
hist(talldata$spin_rate, main = "", col = "royal blue")
boxplot(spin_rate ~ tall_short, mydata, main = "", col = c("light blue","royal
blue"))
title("Spin_rate by Height", outer=TRUE)
skewness(shortdata$spin_rate, na.rm = TRUE)
kurtosis(shortdata$spin_rate, na.rm = TRUE)
skewness(talldata$spin_rate, na.rm = TRUE)
kurtosis(talldata$spin_rate, na.rm = TRUE)
```

```
hist(shortdata$exit_velocity, main = "", col = "light blue")
hist(talldata$exit_velocity, main = "", col = "royal blue")
boxplot(exit_velocity ~ tall_short, mydata, main = "", col = c("light blue","royal
blue"))
title("Exit_velocity by Height", outer=TRUE)
skewness(shortdata$exit_velocity, na.rm = TRUE)
kurtosis(shortdata$exit_velocity, na.rm = TRUE)
skewness(talldata$exit_velocity, na.rm = TRUE)
kurtosis(talldata$exit_velocity, na.rm = TRUE)

hist(shortdata$launch_angle, main = "", col = "light blue")
hist(talldata$launch_angle, main = "", col = "royal blue")
boxplot(launch_angle ~ tall_short, mydata, main = "", col = c("light blue","royal
blue"))
title("Launch_angle by Height", outer=TRUE)
skewness(shortdata$launch_angle, na.rm = TRUE)
kurtosis(shortdata$launch_angle, na.rm = TRUE)
skewness(talldata$launch_angle, na.rm = TRUE)
kurtosis(talldata$launch_angle, na.rm = TRUE)

hist(shortdata$ba, main = "", col = "light blue")
hist(talldata$ba, main = "", col = "royal blue")
boxplot(ba ~ tall_short, mydata, main = "", col = c("light blue","royal blue"))
title("Batting average by Height", outer=TRUE)
skewness(shortdata$ba)
kurtosis(shortdata$ba)
skewness(talldata$ba)
kurtosis(talldata$ba)

hist(shortdata$babip, main = "", col = "light blue")
hist(talldata$babip, main = "", col = "royal blue")
boxplot(babip ~ tall_short, mydata, main = "", col = c("light blue","royal blue"))
title("BABIP by Height", outer=TRUE)
skewness(shortdata$babip)
kurtosis(shortdata$babip)
skewness(talldata$babip)
kurtosis(talldata$babip)

hist(shortdata$slg, main = "", col = "light blue")
hist(talldata$slg, main = "", col = "royal blue")
boxplot(slg ~ tall_short, mydata, main = "", col = c("light blue","royal blue"))
title("Slugging percentage by Height", outer=TRUE)
skewness(shortdata$slg)
kurtosis(shortdata$slg)
skewness(talldata$slg)
kurtosis(talldata$slg)
```

```
hist(shortdata$iso, main = "", col = "light blue")
hist(talldata$iso, main = "", col = "royal blue")
boxplot(iso ~ tall_short, mydata, main = "", col = c("light blue","royal blue"))
title("Isolated power by Height", outer=TRUE)
skewness(shortdata$iso)
kurtosis(shortdata$iso)
skewness(talldata$iso)
kurtosis(talldata$iso)
# At this point data looks okay to move forward with t-tests

par(mfrow = c(1,1))

# Perform individual t-tests to determine if differences exist between tall and
short pitchers
t.test(velocity ~ tall_short, mydata) # t = -5.0923, df = 1424.6, p-value = 4.01e-
07
t.test(effective_speed ~ tall_short, mydata) # t = -4.1286, df = 725.51, p-value =
4.074e-05
t.test(spin_rate ~ tall_short, mydata) # t = -1.6807, df = 722.6, p-value = 0.09326
t.test(exit_velocity ~ tall_short, mydata) # t = -1.0377, df = 733.99, p-value =
0.2997
t.test(launch_angle ~ tall_short, mydata) # t = 0.061186, df = 725.95, p-value =
0.9512
t.test(ba ~ tall_short, mydata) # t = 0.086803, df = 1425.9, p-value = 0.9308
t.test(babip ~ tall_short, mydata) # t = -1.1367, df = 1428.9, p-value = 0.2559
t.test(slg ~ tall_short, mydata) # t = 0.41863, df = 1429.8, p-value = 0.6755
t.test(iso ~ tall_short, mydata) # t = 0.56802, df = 1428.5, p-value = 0.5701

# For the sake of throroughness, checking the non-parametric Wilcoxon rank sum test
# Produced identical results, with only statistically significant differences found
on velocity and effective_speed
wilcox.test(velocity ~ tall_short, mydata) # p-value = 3.089e-07
wilcox.test(effective_speed ~ tall_short, mydata) # p-value = 3.235e-05
wilcox.test(spin_rate ~ tall_short, mydata) # p-value = 0.1114
wilcox.test(exit_velocity ~ tall_short, mydata) # p-value = 0.449
wilcox.test(launch_angle ~ tall_short, mydata) # p-value = 0.712
wilcox.test(ba ~ tall_short, mydata) # p-value = 0.7889
wilcox.test(babip ~ tall_short, mydata) # p-value = 0.2189
wilcox.test(slg ~ tall_short, mydata) # p-value = 0.325
wilcox.test(iso ~ tall_short, mydata) # p-value = 0.1523

# Save datasets for future use
write.csv(mydata, file = "assignment2_data.csv")

# End
```